Course Export - Social Media Research

University

Uni-Potsdam

Semester

WiSe 19/20

---

01 Theory

Explain the Role of Policy-Makers as consumers of research

Policy makers need data to use as the basis for decision making. If they do not have this data they need to rely on intuition, traditions or common sense

Why is it hard to research Privacy ?

Because talk is cheap and people tend to answer with socially desirable answers

How can the reliability of survey's be improved?

- Avoid priming the respondent
- Hide priming topics within questions so they appear less relevant

What are the main types of research in terms of the research process, logic, outcome and purpose ?

There are 4 dimensions.

The purpose can be

- Exploratory: Explore a research topic. Forms the basis of further research
- Descriptive: Identify and classify elements or characteristics of a subject
- Analytical: Extends descriptive research to explain certain findings
- Predictive: Speculate in an intelligent way about the future

The Process can be either Quantitative or Qualitative, the outcome is either basic or applied and the logic is inductive or deductive.

What are the differences in terms of assumptions between **Positivism** and **Interpretivism**

Positivism assumes that a simple and objective reality exists and that knowledge can be obtained in a value free way. Interpretivism assumes that knowledge is just a subjective concept that each researcher perceives.

What can be criticized about positivism as a research paradigm ?

Objective truth / knowledge does not exist, an approach to find it therefore must have its problems. Research methods for positivism also impose certain constraints which might make research harder.

Explain the difference between **Reliability** and **Validity**

Validity describes how well suited for example a question is to gather data about the phenomenon that is supposed to be examined. Such a method is reliable if its result stay consistent over time, between different items/questions and trough the application of different researchers.

Explain the different Methods used in Positivism and Interpretivism

Positivism

- Experiments: Change variables (for example form design) to see how other, dependent, variables react (for example form answers)
- Surveys: Surveys can be deskriptive (describe a certain phenomenon with data) or analytical (show the link between variables)
- Cross-Sectional Studies: Ask the same question(s) in different cultural context's
- Longitude Studies: Gather data over a long amount of time

Interpretivism

- Grounded Theory
- Case Study: Examine a single phenomenon with different methods (interviews, documents, ...) in its natural context
- Ethnography Roadmap: Researcher is embedded in a community and studies its view(s) of the world

What does Triangulation in the context of research methods mean ?

Include

- multiple types of data
- multiple sources of data
- multiple data collection methods

To find a common ground between them all and better understand the research topic.

List all research biases and give examples!

- Selection-Bias: Depending on the selection of the sample results may vary
- Hawthorne-Effect: People behave differently when they are watched
- Interviewer-Bias: Interviewers, trough their behavior/use of methods/questions, may introduce a bias
- Social-Desirability-Bias: People tend to answer in a way they think is socially desirable
- Non-Recall-Bias: People forget things or recall them differently
- Non-Response-Bias: Trough Participants not responding the sample might become unrepresentative
- Interpretation Bias: Researchers might introduce their bias while interpreting the data

What is literature review and why is it needed ?

Literature review is the **summary** of a specific question. It is needed in order to

- get background knowledge about a topic and synthesize prior findings
- find out where research **exists** and where research is **lacking**
- as a foundation for theory building

What determines the quality of a Literature Review?

There are 4 directions, a Literature Review must

- have appropriate breath and depth
- be performed with rigor and consistency
- be clearly formulated and have sufficient brevity
- be an effective analysis and synthesis

How does Research in an area develop ?

There are 4 steps in emergent research fields

- Ad-Hoc: Simple categories are formed from empirical data
- Taxonomies: Links between categories are established
- Conceptual Frameworks give propositions on summaries and explanations in a field
- Theoretical Systems establish laws that are contained in Theories

What is Keyword search, in which of the 3-Stage-Research-Process is it used and what are its drawbacks ?

Keyword search is used in the first step of a Literature Review (Identifying relevant literature).

Journals, Databases and Conferences are searched for Keywords to find relevant literature.

A limitation is that keywords have limited time validity and words people use might change over time, thus forward (who cited this paper?) and backward (what papers where cited in this paper) search need to be performed in order to find all relevant literature.

Explain the differences between Concept-Centric and Author Centric literature analysis!

Both, the analysis of the literature and the writing of the analysis, can be performed in author or concept centric ways.

When working **author** centric sources are listed by paper/author and concepts presented in an article are just listed. The analysis is formulated in an author centric way stating which author presented certain concepts:

> **Example**: Krasnova (2020) reports that Social Media Research is important.

When working **concept** centric a table might list concepts and all authors who wrote about it. The analysis is formulated based on the concept and differences are shown when authors disagree

> **Example**: Studies report that SMR is important (Krasnova 2020) but other researchers disagree and state that it isn't (Gronau 1994)

Explain what a Concept Matrix and an augmented Concept Matrix with units of analysis are ? (Context of Concept vs Author centric Literature Analysis)

A Concept Matrix is a table containing concepts as columns and articles as rows. Each cell contains a value that indicates if the concept is discussed in the matching article.

An augmented concept matrix with units of analysis uses categories to further differentiate concepts or articles. For Example a concept about social media use can be split into papers discussing it in a private or a corporate environment.

Explain the differences between Concepts, Constructs and Variables

- Concept: A term to express and idea
- Construct: A concepts that can be used to measure a not directly observable variable (e.g. Self-Esteem, IQ)
- Variable: Something that can be directly measured (e.g. Temperature, Weight)

What is a Theory ?

A Theory provides **explanations of an observed phenomenon**. It also must answer the **why** question (explains why explanation holds).
It has three critical components:

- Proposition or Hypothethis
- An Explanation of the phenomenon
- A representation of a systematic view of phenomena by noting the relations between certain constructs or variables

More Info Session 3 Slide 44

Why is Qualitative Research important ?

Qualitative research is important to obtain a deep, bottom-up understanding of a research topic which often cannot be obtained in any other way.

Explain differences between Qualitative and Quantitative Research!

Qualitative	Quantitative
in depth understanding	generalizable results
close to the reserach subject	distant from research subject
explore emergent topics	confirmation of theories
rich, deep data from small sample	hard, reliable data from big samples

Explain the main categories of Qualitative Research Procedures

Qualitative Research can be either

- Direct: The participants know the topic/Research question (e.g. Focus Groups)
- Indirect: Trough the use of projection techniques the research topics stays hidden for the participants

Explain the procedures and positive sides of Focus Groups!

A group of 8-10 participants holds a moderated discussion. Moderators are prepared and know the topic well.

They **start** with the rules of the discussion and pose an opening question. They then moderate the discussion and guide it trough the use of prepared questions. In the **end** they summarize and ask for confirmation of their summary. Everything is recorded and transcribed and later analysed using qualitative methods.

Focus groups are good to promote self disclosure and the discussion of feelings and beliefs. They are also good to get new an innovative ideas.

Explain the types of Interviews ?

Interviews can be

- structured: all questions are pre-formulated and cannot be changed
- semi-structured: there are questions that guide the subject but there are no hard rules for the conversation
- unstructured: questions might exist but the interview is mainly a conversation

What types of probes can be used in interviews ?

There are probes for

- clarity: give an example ? explain again ?
- relevance: how does this relate to the issue ?
- dimension: are there other perspectives/opinions ?
- significance: how does this affect you ? what is most important ?
- comparison: was there a situation where xyz did not happen ?
- bias: why do you think this way ?

What are limitations of In-Depth Interviews ?

- subjects might be dishonest
- priming questions can influence the result
- the interviewer influences the interview trough for example his gender or race

Explain the essentials of the Delphi-Method

A global panel of experts is questioned trough an anonymous questionnaire. This is done in multiple rounds with improved questions.

This avoids confrontation between experts and gives a lot of depth trough anonymity.

What types of projektive techniques exist ?

- Word association: participants are asked to say a word they associate with another word
- Sentence/Paragraph/Story completion: participants are asked to finish a text
- Interpretation: participants are asked to interpret the behavior of others

Tasks from other topics might be included within the research tasks in order to hide the true purpose of the research.

What are pros and cons of Projektive Techniques ?

Pros

- allows discovery of sensitiv information and discussion of undesirable topics
- can be an ice breaker in focus groups

Cons

- extremely subjective in interpretation and therefor unreliable
- experts needed for interpretation and therefor expensive

What are the limitations of Focus Groups (5 M's)?

Limitations are given by the 5-M's

- Misuse: Researchers might think the results of a focus group are conclusive while in reality they can only be used as exploratory data
- Misjudge: When interpreting focus groups a subjective bias might lead to wrong conclusions
- Moderation: The quality strongly depends on the moderator
- Messy: Data is messy and hard to analyze
- Misrepresentation: Results are not representative

Explain the essentials of the Delphi Method.

1. Exercise in group communication among a panel of geografically dispersed experts (allows experts to deal systematically with a complex problem, aims to obtain a reliable response to a problem from experts)
2. Experts do not interact with each other (anonynomity, controlled feedback, statistical response)

What are the major steps of the Delphi Method?

1. Selection of expert panel(s)
2. Development of the first round questionnaires (structured) & testing of the questionnaire
3. Transmission to the panellists
4. Analysis of 1st reponses
5. Preparation of the 2nd round (proving the answers from the first round)
6. Transmission of the second round
7. Analysis of the second round

To be continued if needed

What is the outcome of the Delphi Method? Which problem occurs?

Outcome -> nothing but opinions of experts

Problem: the results of the sequence are only as valid as the opinions of the experts who made upt the panel

Regarding the interpretation bias, which qulitative data collection technique is most useful for collection unbiased data?

Focus Groups --> group is observed by scientific team and if an "uncritical" topic is picked, participants are very likely to tell the truth

Which qualitative data collection method is the best to collect innovative information? Why?

Focus Group

Why?

1. Snowballing -> Comment of one person triggers a chain reaction from others.
2. Stimulation -> Excitement over ideas increases in groups, so people want to express ideas.
3. Serendipity -> Ideas more likely to arise accidentally in groups.

Why is sexuality not a suitable topic for focus groups? Name the caracteristics for a suitable topic.

People might be devided or angry and the topic involves intimate details.

A suitable topic:

- where people have sth to share
- where diversity of viewpoints is of advantage
- where discussion can be easily triggered
- where to goal is to get insights into human motivations and behavior

What types of codes exists?

Coding can be based on pre set (theory-driven) or emergent (data-driven) categories

Codes can be manifest (directly observable, e. g. times a respondent talks about a specific topic) or latent (not directly observable, e. g. information overload)

Name different methods for primary and secondary qualitative data gathering!

Primary

- Interviews
- Focus Groups
- Observations
- Diaries

Secondary

- Reports
- Articles
- Broadcasts/Films
- Stories/Documents

What are the steps of Content Analysis!

1. Familiarize yourself with the data, check its usability and depth

2. Focus on which questions can be answered by the data
3. Categorize/Code Data
4. Identify patterns and form bigger categories, interpret frequency, co-occurrence patterns and sequence of occurrence
5. Interpret your findings

Explain the difference between **Thematic Coding and **Content Analysis**!**

Thematic coding is every form of categorization qualitative data. Content Analysis is a very rigorous and reliable form of Thematic Coding.

Explain the different types/properties of Codes in Content Analysis!

Codes can either be pre-determined based on what the researcher expects or they can emerge during analysis.

Codes can also be assigned to the exact wording of an answer (manifest) or to the underlying meaning of the answer (latent).

Lastly it can be decided if each section can receive one or multiple codes.

Explain the Process/Steps of Coding!

1. Immerse yourself into the data trough reading it thoroughly
2. Generate a first set of codes
3. Apply, refine, elaborate, split or group codes
4. Test codes for reliability (trough comparison with other coders)

Explain the concept of Inter-Rater Reliability!

Inter Rater Reliability (IRR) describes the extend to which multiple coders come to the same conclusion when coding data.

There a different Methods to measure IRR

- Percentual Agreement: easy, widely used but apparently not sufficient/perfect
- Other methods include: Holsti's method, Scott's pi, Kippendorfs alpha and Cohen's Kappa

Explain the Units of Analysis in the context Content Analysis!

Determines what unit is analyzed. For example it can be Phots, Interviews, Words, Blogs, Movies, Concepts. This depends strongly on the purpose of the research.

What are the components of a Codebook for Content Analysis ?

Codebooks are typically tables that contain the following information for each code

- a label or name
- a definition of what the code means
- a description on how to decide if the code should be applied
- qualifications, elaborations and/or exclusions
- positive and negative examples

Explain Frequency Analysis!

Analyzing the frequency of certain codes/categories appearing in a data set. Importance of codes is then determined by the frequency of its occurrence.

Shortly summarize the history of Grounded Theory!

Glaser and Strauss developed it as a tool for their research. They later split directions with Strauss describing procedures that Glaser sees as a completely different method. Glaser makes a different proposal. In the end nobody really knows what the differences are and how to really define Grounded Theory.

What are the main differences between Glaser's and Strauss's Grounded Theory ?

Strauss uses one coding paradigm while Glaser proposes 18 families of codes because he things Strauss's method is to restrictive.

Strauss allows for prior knowledge to influence the GT while Glaser expects an unbiased "clean-slate" research approach.

What are the key Advantages of Grounded Theory as opposed to other qualitative methods ?

The are clear rules on where to **start** how to **perform** the research and when to **end** the process.

What are the critical Features of Grounded Theory ?

1. Simultaneous collection and analysis of data
2. Creation of codes from data and **not** from pre-defined categories (emergent codes)
3. Allows the discovery of basic social processes
4. Inductive construction of abstract categories
5. Theoretical sampling to refine categories
6. Usage of analytical memos between coding and writing
7. Integration of categories into a theoretical framework

How does Data Collection work in Grounded Theory Research ?

Both primary and secondary data collection methods can be used if applicable for the research question. Questions and methods can be modified during the research when theories emerge.

Explain what Constant **Comparison means in the context of **Grounded Theory** ?**

Constant comparison between the data from multiple categories. Allows to understand what is happening, under which conditions things happen and what data means.

Pretty abstract on the slides, further definition would be useful.

Explain what **Theoretical Sensitivity means in the context of **Grounded Theory**!**

Theoretical sensitivity develops as researchers get increasingly familiar with the data they analyse.

Researcher should not try to fit data to previously found categories and use categories and codes that emerge during analysis.

Past research should only be used to inform decisions. Concept of "Open mind but not empty head".

Explain **Theoretical Sampling in the context of **Grounded Theory****

Theoretical Sampling is the process of researchers deciding what data to collect next. They usually start with a random sample and then choose based on the data they have already gathered.

Explain Open, Axial and Selective Coding for Grounded Theory!

- Open Coding: label/code meaningful parts of the data, build basic categories of related codes
- Axial Coding: develop "mini-theories" on how categories relate and determine the "Core Category"
- Selective Coding: Formalize categories belonging to the Core Category into theoretical Frameworks

What is the "Core Category" in Grounded Theory ?

The Core Category is a central category that has many relations to other categories. Emergent theories normally center around it.

Choosing a good name is important so that people understand what is meant.

Describe/Explain the Coding Paradigm by Strauss ?

src="https://i.imgur.com/iKf4E1Z.png">

What is the Role of Memos in Grounded Theory ?

Memos are an intermediate between coding and the writing of the first draft.

They document the development process of theories and include first ideas and hypothesis. They are used and modified during analysis.

What is **Theoretical Saturation in the context of **Grounded Theory****

Theoretical Saturation is reached when no additional data is found in new rounds of data gathering.

Here a trade off has to be made between the completeness and the cost to continue research. Good Theoretical Sampling lowers the chance of missing data.

What are typical Outcomes of Grounded Theory ?

- New Theories
- Models: Definitions of abstract variables and their relations
- Rich Descriptions: Narratives of empirical observations without any abstraction

Why/When should Participatory Observations be used as a means of data collection ?

- Enables researcher to observe the life of an "insider" while staying an "outsider"

- Create a "written photograph" of a situation
- Provide a nuanced understanding of a situation
- Uncover previously unknown concepts

What are benefits and drawbacks of Participatory Observations ?

Pros

- Remove the subjective factor of reporting
- Give deep insights into phenomenons/social context
- Can uncover previously unknown concepts

Cons

- Very time consuming
- Hard to document findings while observing
- Very subjective as what someone sees is different from what is interpreted
- Personal (ethical, racial, social, ..) background can influence observations

What are the typical steps of Participatory Observation ?

- Find a location, analyze it and determine the best time to do the observation
- Observe and take Notes about everything
- Document in depth directly after the observation finished
- Analysis

What are the 3 steps in the `3-Stage-Process` of research ?

1. Identifying Relevant Literature
2. Literature Analysis
3. Theory Building

02 R Data Analysis

What are the different Data Types in R

- Character: Single character or string
- Numeric: All types of numbers (including floating point). Automatically converted by R to appropriate sub-types.
- Integer: Full Integers, no automatic conversion
- Logical: TRUE/FALSE
- Complex: Allows the use of complex arithmetic / imaginary numbers e.g. `2+3i`
- Raw: Stores data as raw bytes. Displayed as HEX when printed.
- Function: Callable R functions.

Explain the basic syntax of R Markdown ?

R Markdown are Markdown based documents that can include executable R Code. Code can either be Inline starting with `` `r sum(1:5)` `` or as a Code Block starting with

```
```{r block_name, include=TRUE, echo=TRUE, message=TRUE, warning=FALSE}
code

```
```

R Markdown documents are usually exported as PDF, Word or HTML files.

What is the syntax to create a matrix with 2 columns of data?

```
> foo = matrix(1:10, ncol=2)
> foo
     [,1] [,2]
[1,]    1    6
[2,]    2    7
[3,]    3    8
[4,]    4    9
[5,]    5   10
```

Explain the difference between an Array and a Matrix!

An array is an n-dimensional object while a matrix has only two dimensions (rows/columns)

What is the difference between a Matrix and a Data Frame

A matrix can only contain one type of data while Data Frames can contain multiple types.

What is Vector recycling ?

Vector recycling is a concept in R where, if two vectors of unequal length are part auf an operation, the elements of the shorter vector repeat until they match the length of the longer vector.

```
> v1 = c(0,0,0,0)
> v2 = 1:2
> v2
> v1 + v2
[1] 1 2 1 2
```

What is lazy evaluation in R?

Lazy evaluation means that a symbol is only evaluated when it is actually needed.

What is the syntax to install an package in R?

```
install.packages("ggplot2")
```

Describe how Variable assignment works in R!

Variables can be assigned using either the = or the `<-` (and `->`) operator.

Multi assignments like `a = b = c = 1` are also possible (in this case a,b,c are equal to 1)

What does Element Wise Execution refer to?

Element wise execution means that R executes an operation on an object for each of its sub objects.

```
> v1 = 1:10
> v1
[1]  1  2  3  4  5  6  7  8  9 10
> v1 + 5
[1]  6  7  8  9 10 11 12 13 14 15
```

What types of special values (e.g. missing values) exist in R and how are they shown ?

- `NaN` - Not a Number
- `Inf` - Infinity
- `NA` - Not available (missing data), can be checked with `is.na(foo)`

What is transpose and how does it work in R?

Transpose switches columns and rows of a matrix. It is performed using the `t()` function.

```
> test
      [,1] [,2]
 [1,]    1    1
 [2,]    2    2
 [3,]    3    3
 [4,]    4    4
 [5,]    5    5
 [6,]    6    6
 [7,]    7    7
 [8,]    8    8
 [9,]    9    9
[10,]   10   10
> t(test)
     [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8] [,9] [,10]
[1,]    1    2    3    4    5    6    7    8    9    10
[2,]    1    2    3    4    5    6    7    8    9    10
```

How do you access the element in the 2nd column and 4th row of a matrix named M?

`M[4,2]` or generally `matrix[rownum, colnum]`

Explain the build in function `apply()`

`apply(matrix, row_or_col, fun)` applies a function (e.g. `sum`) to either the rows or columns of a matrix.

Example function `sum`: If `row_or_col` is 1 then all cells in a row are added together and a sum is displayed for every row. If it is 2 then all columns are summed.

```
> m
     [,1] [,2]
[1,]    1    1
[2,]    2    2
[3,]    3    3
[4,]    4    4
[5,]    5    5
[6,]    6    6
[7,]    7    7
[8,]    8    8
[9,]    9    9
[10,]   10   10
> apply(m,1,sum)
[1]  2  4  6  8 10 12 14 16 18 20
> apply(m,2,sum)
[1] 55 55
```

What do the operators `%%` and `%/%` do ?

- `3 %% 2` - calculates `3 modulo 2 == 1`
- `4.2 %/% 2` - integer devision cutting of decimals from the result

How do you check if a variable `m` is a matrix ?

```
> class(m)
[1] "matrix"
```

What does the `which()` function do ?

Returns the index of all elements in a vector or array matching the condition

```
> v1
[1]  1  2  3  4  5  6  7  8  9 10
> which(v1 > 5)
[1]  6  7  8  9 10
```

Explain the different data Structures: Vector, Matrix, Array, List and Data Frame

The following can always only contain one type of data:

- Vector: 1-Dimensional - basically a sequence of objects
- Matrix: 2-Dimensional - basically a table with rows and columns
- Array: n-Dimensional - a multi dimensional table with each cell having the possibility to contain new Arrays

For mixed data types

- List: 1-Dimensional
- Data Frame - 2-Dimensional

How can rows and columns be added to Data Frame ?

- rows `rbind(frame, list(one, value, for, each, column)`
- columns `cbind(frame, colname=c(one, value, for, each, row))` or `frame$colname = c(one, for, each, row)`

How to concatenate Strings in R ?

```
> paste("foo","bar")
[1] "foo bar"
```

How do you get part of a string in R ?

use the function `substr(string, start_index, end_index)`

```
> substr("foobar", 1,3)
[1] "foo"
```

How do you get an overview of a Data Frame ?

```
> summary(df)
      X1              X2
Min.   : 1.00   Min.   : 1.00
1st Qu.: 3.25   1st Qu.: 3.25
Median : 5.50   Median : 5.50
Mean   : 5.50   Mean   : 5.50
3rd Qu.: 7.75   3rd Qu.: 7.75
Max.   :10.00   Max.   :10.00
```

If `df` is a Data Frame, what does `sum(is.na(df))` do ?

Returns the number of missing values.

What can the `table()` function be used for ?

`table()` is useful for calculating the (conditional) frequency of data in a list (or anything else that can be interpreted as factors).

```
> df
    type
1    male
2  female
3   other
4    male
5    male
6  female
7   other
8  female
9  female
10  other
> table(df$type)

female   male  other
    4      3      3

> table(df$type != 'male')

FALSE  TRUE
   3     7
```

What does `tapply(FB.df$likes, FB.df$gender, mean)` do, if `FB.df` is a Data Frame containing Facebook data with gender and likes?

Returns the mean like count for each gender.

Explanation: `tapply()` applies the function `mean` to the objects in `FB.df$likes` treating the second parameter `FB.df$gender, mean)` as factors to group by.

What does the following function return ``` library(rtweet) x = search_tweets("social media reserach", n = 18000, include_rts = FALSE, retryonratelimit = TRUE) ```

A Data Frame containing all tweets found with the search term "social media research" in the last few days with a maximum of 18.000 entries excluding retweets.

What do the functions `head()` and `tail()` do ?

Return the first (head) or last (tail) 6 rows of a dataset

If `timeline` is a Data Frame that contains a column `favorite_count` what does `timeline <- timeline[order(-favorite_count), ]` do ?

Orders the timeline Data Frame by the number contained in `favorite_count` in descending order.

create a vector with the numbers 15 to 20

without variable storage a) c(15:20) b) 15:20 c) c(15,16,17,18,19,20)

with storage a) v <- c(15:20)

usw.

We have the vector: name <- c("Steve, "Sophia", "Oliver", Harry", "Mark") How do you access Sophia?

name[2]

name <- c("Steve","Sophia","Oliver","Harry","Peter","Olivia","Marc") how do you acces the 2. and 4. name?

name [c(2,4)]

name <- c("Steve","Sophia","Oliver","Harry","Peter","Olivia","Marc") determine the names of all users except the 2. and the 4.

name[c(-2, -4)]

or

name[c(1,3,5,6,7)]

age <- c(15,16,17,18,18,19) determine those older than 16

age[age>16]

``` name<- c("Lilly", "Benedict", "Daniela") age <- c(15,16,17) ``` add the vector of names to the age, as a second column.

```
> cbind(age,name)
     age  name
[1,] "15" "Lilly"
[2,] "16" "Benedict"
[3,] "17" "Daniela"
```

create a matrix with 7 rows

```
m <- matrix(nrow = 7)
m

Output:

        [,1]
[1,]   NA
[2,]   NA
[3,]   NA
[4,]   NA
[5,]   NA
[6,]   NA
[7,]   NA
```

agenames <- c(16,16,19, "Johanna", "Mark", "Sophia") create a matrix with this vector, where age and names are in 2 seperate columns.

matrix(agenames, ncol =2)

What is the output of 9 %%3?

0, because %% is the modulo operator

What is the output of 9.6 %/% 3?

3 because %/% is integer division. Numbers will be rounded.

What is the output of abs(-4.4)?

4.4 because abs() means absolute value ("Betrag" in German)

x <- "abcdefg" substract "cde"

substr(x, 3, 5)

a <- "hello" b <- "Olga" Concatenate these 2 characters.

paste(a,b, sep="")

x <- c(1:10) get the average of x

mean(x)

what is the output of mean(c(1,2,NA), na.rm =TRUE) ?

1.5 because if the parameter na.rm = TRUE, then the missing values will be disregarded.

How can you create a vector as a sequence of numbers and specify starting point, end point and steps in between?

v19 <- seq(from=5, to=300, by= 4)

How can you sort an unsorted vector (v) ?

sort(v)

numeric values are sorted cardinally

Check if there are any missing values in the vector (v).

anyNA(v)

What means "coercion" in R and what are the rules?

Converting objects from one class to another is called *coercion*.

Rules:

numeric+character= character

xx<-c(1,"hi")

logical+character= character

yy<-c(TRUE, "hi")

numeric+logical=numeric

zz<-c(1, TRUE)

How can you change the class of an object in R? (forced coercion)

as.<desired class>()

e.g.

as.numeric("1")

as.character(1:3)

What means "indexing" ?

Indexing means selecting a subset of the elements in order to use them in further analysis.

R uses square brackets [] to specify the position of values to be extracted. A standard way to index is to specify the position of an element by a number or with help of logical operators.

e.g.

v[2:3]

v[c(1,3)]

v[v > 11 & v < 14] # logical indexing

m[,]

`list1<-list(1,"a", TRUE, c(2.2, 5.1))` 1. Return the 4th element of list1. 2. Return the second value of the 4th element of list1.

1. list1[[4]]
2. list1[[4]][2]

Lists can be indexed in two ways:

- `[i]` returns a list of elements. i can be an integer or a vector `[i:j]`, just like when indexing a vector

- `[[i]]` returns a single element. i can only be a single value

Data frame (dat): Replace the first element in column "y" with 20.

dat[1,"y"]<-20

Delete a column (x) in the data frame (dat).

dat$x <- NULL

dat <- dat[,-1]

What are special keywords that you cannot use in R?

break, for, Inf, next, TRUE, else, function, NA, repeat, while, FALSE, if, NaN, return

What is a function? Why do we need operators?

A set of statements organized together to perform a specific task. Operators are symbols used as a function (e. g. +, -)

What is the difference between a bar chart and a histogram?

A bar chart represents data in rectangular bars with length of the bar proportional to the value of the variable.

A histogram represents the frequencies of values of a variable bucketed into ranges.

Main difference: Histogram groups the values into continuous ranges.

Give the mean value of 1,2,3,4, NA.

mean(c(1,2,3,4, NA), na.rm=TRUE)

This function [mean (x, na.rm=TRUE)] removes all missing values. If you would just enter mean(c(1,2,3,4,NA)) the output would be NA.

What is the function to create a bar chart in R?

barplot (x)

What is the most important package for data scraping with APIs ?

library (rtweet)

Show the range of the numbers 2,9,10,13. What will be the output? Additionally print the difference between the largest and the smallest value.

y <- c(2,9,10,13)

range (y)

Output: 2 13

Showing the difference: diff(range(y))

How to search for tweets by hashtag?

rt <- search_tweets (q ="#rstats", n = 3000, include_rts= FALSE)

q = name of the hashtag

n = max. number of tweets in your search

include_rts = include retweets TRUE or FALSE

What is cumulative summing? What will be the output of the following code: `cumsum(1:3)`

for cumulative summation, the first value is output and then added to the subsequent one. The result is then added to the next value. In this specific case:

```
> cumsum(1:3)
[1] 1 3 6
```

How to search for tweets by user?

Gates <- get_timeline (“@BillGates”, n = 1000)

What does the which function?

Returns the position of the elements in a logical vector which are TRUE. e. g. x <- c(3,2,1) // which (x==3); Output: 1

Search a number for users, who used a certain hashtag in their profile bios.

usrs <- search_users (“#rstats”, n=1000)

Search 3000 most recently favorited statuses by a user.

jkr <- get_favorites (“clindner”, n=3000)

Discover what´s currently trending in San Francisco on Twitter

sf <- get_trends (“san francisco”)

Describe the analytics workflow.

Data Access

Data processing and normalization

Data analysis

Insights

How to identify and replace items from tweets?

gsub (pattern, replacement, x)

DFname$text <- gsub (“http.*”, ””, DFname$text)

What is a word cloud?

A word cloud is a visual representation showing the frequency of words in a document by varying the size of words.

What is a comparison cloud and how to use it?

A comparison cloud compares the relative frequency with which a term was used in two or more documents in a visualization.

comparison.cloud(tdm, random.order=FALSE, colors = c("indianred3","lightsteelblue3"),title.size=2.5, max.words=400)

What is a commonality cloud and how to use it?

The commonality cloud is the complement to the comparison cloud. It shows only those words that appear in all documents and their combined frequency across documents. A commonality cloud is useful for showing the amount of conceptual overlap between two documents.

commonality.cloud(tdm, random.order=FALSE, scale=c(5, .5),colors = brewer.pal(4, "Dark2"), max.words=400)

What are looping constructs?

for, while, repeat

What is the apply (...) function?

apply() is a R function which enables to make quick operations on matrix, vector or array. The operations can be done on the lines, the columns or even both of them.

Example: Execute a given function to the rows (index 1) or column (index 2)of a matrix or an array. [apply(x,2,sum].

Which function can be used to concatenate strings?

paste (...), e. g. paste ("x", 1:3, Sep="M"), Output: "xM1" "xM2" "xM3"; Important: Sep = M, puts a "M" between the elements that are concatenated by the paste function

Split the following string: hello

strsplit("hello", ""), Output: h e l l o

Round x = 3.3456 to the number of two decimal places

round(3.3456, digits=2); Output: 3.35

How to retain only unique rows from an input tbl (e.g. results from search_tweets)?

Use the distinct() function.

searchresults1 <- distinct(search_tweets(q= wordlist1, n=10, include_rts= FALSE))

Which functions can be used to transform letters to lower or upper case?

toupper("hello"); Output: HELLO

tolower("Hello"); Output: hello

How to determine only tweets of a certain language (en) in your search results?

searchresults1en <- searchresults1[which(searchresults1$lang=="en"),]

What is the function of aggregate? Write down the code to aggregate the mean price of products from a list, where products are given with detailed brand and price.

Aggregate "aggregates" the inputted data by applying a special function (FUN = (...)), e. g. the mean value. The function is applied to each column of the sub-data.frame defined by the input parameter: aggregate (Object, by=list(Input Parameter), FUN=Function); e. g. `aggregate(products$price, by=list(products$name), FUN=mean)`

How can you add more columns of a data.frame to the aggregate function?

Using more by parameters, e. g. `aggregate(products$price, by=list(products$name, products$brand), FUN=mean)`. Now products are separated by brands and the mean value will be calculated by each product and brand, e. g. Jeans Calvin Klein, Jeans Esprit, Jeans Pepe Jeans

How can you clean your Twitter search results from possible stop words?

```
data ("stop_words")

twitterdata_clean <- twitterdata %>% anti_join(stop_words)
```